9. Approximate inference by sampling

- Markov Chain Monte Carlo methods
- Metropolis-Hastings algorithm
- Gibbs sampling
- Bounding mixing time via spectral analysis
- Bounding mixing time via coupling

Approximate inference with samples

- inference problem in graphical model

$$\mu(x) = \frac{1}{Z} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j)$$

- belief propagation
  - fast (especially on sparse graphs) and very popular
  - deterministic
  - computes (approximation of the) marginals
- approximate inference with samples
  given samples $\{x^{(1)}, \cdots, x^{(N)}\}$ from distribution $\mu(x)$

$$\frac{1}{N} \sum_{j=1}^{N} \mathbb{I}(x_i^{(j)} = x_i) \to \mu(x_i)$$

  gives an approximate marginal
  - slower and difficult to decide when to stop
  - randomized

Generating samples from a distribution

generating samples from $\mu(x)$	generating samples from $\mu(x_i)$
Markov Chain Monte Carlo methods	sequential Monte Carlo methods
Metropolis-Hastings algorithm	particle filtering

- **Markov Chain Monte Carlo methods** work as follows
  - construct a Markov chain P whose stationary distribution is equal to μ
  - start from an arbitrary realization $x^{(0)}$ and run the Markov chain until it converges to its stationary distribution
  - this gives a sample from $\mu(x)$
- how do we construct such a Markov chain P?
- how long does it take for the Markov chain to converge?

Metropolis-Hastings algorithm

- Markov chain with a finite state space
  - a Markov chain is defined by a state space $\mathcal{X}^n$ and a $|\mathcal{X}|^n \times |\mathcal{X}|^n$ dimensional transition matrix P such that

$$P_{xy} = \mathbb{P}(x_{t+1} = y | x_t = x)$$

  - stationary distribution of a Markov chain is a $|\mathcal{X}|^n$-dim row vector of distribution such that

$$\pi^T P = \pi^T$$

  - a Markov chain is **reversible** if there exists a probability distribution π such that the **detailed balance equation** is satisfied:

$$\pi_x P_{xy} = \pi_y P_{yx} \qquad \text{for all } x, y$$

  - further, the corresponding π is a stationary distribution

$$(\pi^T P)_x \;=\; \sum_y \pi_y P_{yx} \;=\; \sum_y \pi_x P_{xy} \;=\; \pi_x$$

- the strategy is to construct a Markov chain P such that it is reversible, so that we can apply spectral analysis techniques, and has the desired stationary distribution $\pi_x = \mu(x)$

- **Metropolis-Hastings algorithm**
  - start with a candidate transition matrix K, which we will modify to create P
  - to ensure unique stationary distribution, it is sufficient to have
    - ⋆ $K_{xx} > 0$ for all $x \in \mathcal{X}^n$, and [aperiodic]
    - ⋆ the undirected graph $G(K) = (\mathcal{X}^n, E(K))$ is connected, where $E(K) \equiv \{(x, y) : K_{xy}K_{yx} > 0\}$ [irreducible]
  - we want the transition matrix to satisfy the detailed balance equation with μ, but instead for each pair (x, y), suppose the following holds without loss of generality, i.e. instead of $\mu(x)K_{xy} = \mu(y)K_{yx}$ we have

$$\mu(x)K_{xy} \;>\; \mu(y)K_{yx}$$

  - the trick is to remove some 'probability mass' from the larger one
    - ⋆ define $R_{xy} \equiv \min\left(1, \frac{\mu(y)K_{yx}}{\mu(x)K_{xy}}\right)$
    - ⋆ let

$$P_{xy} \quad \equiv \quad \begin{cases} K_{xy}R_{xy} & \text{if } y \neq x \\ 1 - \sum_{y \neq x} P_{xy} & \text{if } y = x \end{cases}$$

    - ⋆ then, P satisfies the detailed balance equations w.r.t μ, and hence μ is a stationary distribution of P

$$\mu(x)K_{xy}R_{xy} = \mu(x)K_{xy} = \mu(x)K_{xy}\frac{\mu(y)K_{yx}}{\mu(y)K_{yx}} = \mu(y)K_{yx}R_{yx}$$

- challenges with **Metropolis-Hastings algorithm**
  - do we need μ to construct P?
    we only need $\frac{\mu(x)}{\mu(y)} = \prod_{(i,j)\in E} \frac{\psi_{ij}(x_i,x_j)}{\psi_{ij}(y_i,y_j)}$
    which can be evaluated efficiently. In particular, we do not need to compute the partition function Z.
  - how do we store K and P with dimensions $|\mathcal{X}|^n \times |\mathcal{X}|^n$?
    consider this construction as describing a sampling process
    - at time t generate a candidate sample x' according to $K(x^{(t)}, x')$, which possibly has a simple structure
    - *accept* the candidate state with probability $R_{x^{(t)},x'}$
    - otherwise *reject* and keep current state
- **theorem.** Metropolis-Hastings algorithm finds ℓ_1-projection of K onto the space of reversible Markov chains with stationary distribution μ

$$P \;=\; \min_{Q\in R(\mu)} \sum_x \sum_{y\neq x} |\mu(x)K_{xy} - \mu(x)Q_{xy}|$$

- the 'art' is in choosing appropriate K, since bad choice of K results in a Markov chain with slower convergence
- if 'spread' is too narrow, we are not exploring
- if 'spread' is too large, acceptance rate can be low
- **example.**

$$K = \frac{1}{|\mathcal{X}|^n} \mathbf{1}\mathbf{1}^T, \qquad R_{xy} = \min\Big(1, \prod_{(i,j)\in E} \frac{\psi_{ij}(y_i, y_j)}{\psi_{ij}(x_i, x_j)}\Big)$$

all pairs are sampled with equal probability (as per K), but many of them might be unlikely and be rejected with high probability

Gibbs sampling

- **Gibbs sampling** defines P_{xy} as
  - at each time step, first select $i \in \{1, \ldots, n\}$ from a uniform distribution
  - set $y_{[n]\setminus i} = x^{(t)}_{[n]\setminus i}$ and sample y_i from $\mu(y_i | x_{[n]\setminus i})$
- for sparse graphs, it is easy to evaluate $\mu(y_i | x_{[n]\setminus i}) \propto \prod_{j \in \partial i} \psi_{ij}(y_i, x_j)$
- thus generated P satisfy the detailed balance with μ
  - suppose x and y only differ in exactly one position i

$$
\begin{aligned}
\mu(x) P_{xy} &= \mu(x) \frac{1}{n} \mu(y_i | x_{[n]\setminus i}) \\
&= \mu(x_i | x_{[n]\setminus i}) \mu(x_{[n]\setminus i}) \frac{1}{n} \mu(y_i | x_{[n]\setminus i}) \\
&= \underbrace{\mu(x_{[n]\setminus i}) \mu(y_i | x_{[n]\setminus i})}_{\mu(y)} \underbrace{\frac{1}{n} \mu(x_i | x_{[n]\setminus i})}_{P_{yx}}
\end{aligned}
$$

  - otherwise, $P_{xy} = 0$ if x and y differ in more than one position
- the resulting dynamics of the Markov chain is called **Glauber dynamics**

- Gibbs sampling and the analysis of Glauber dynamics is used in
  - Noisy best response in coordination games
    [L. Blume, Games Econ. Behav., 1995]
  - Learning Boltzmann machines (*contrastive divergence*)
    [G. Hinton, Neural Computation, 2002]
  - ...

Mixing time

- two common ways to analyze the mixing time of a (reversible) Markov chain is **spectral analysis** and **coupling**
- **Define.** **ϵ-mixing time of** P is the smallest time such that for all $t > T_{\rm mix}(\epsilon)$

$$|(p^{(0)})^T P^t - \pi^T|_{\rm TV} \quad \leq \quad \epsilon$$

for any initial distribution $p^{(0)}$, where $|x-y|_{\rm TV} = \sum_i |x_i - y_i|$ is the total variation distance

- **Theorem.** we can show that $|(p^{(0)})^T P^t - \pi^T|_{\rm TV} \leq |\lambda_2|^t \Big(\frac{1}{\sqrt{\pi_{\min}}}\Big)$, where $|\lambda_2| < 1$ is the second largest eigenvalue of P
this implies

$$T_{\rm mix}(\epsilon) \quad \leq \quad \frac{\log \frac{1}{\epsilon\sqrt{\pi_{\min}}}}{\log(1/|\lambda_2|)} \quad \leq \quad \frac{\log \frac{1}{\epsilon\sqrt{\pi_{\min}}}}{\underbrace{1-|\lambda_2|}_{\text{spectral gap of } P}}$$

- $\frac{1}{1-|\lambda_2|}$ is called the *relaxation time* of a Markov chain

- spectral properties of Markov chains

Property 1. $\pi P = \pi$ and $P\mathbb{1} = \mathbb{1}$ corresponding to $\lambda_1 = 1$

Property 2. $\pi^T = \pi^T P = \cdots = \pi^T P^t$

- spectral properties of reversible Markov chains

Property 3. $P = \Pi^{-1/2} S \Pi^{1/2}$ for some symmetric matrix S and $\Pi = \text{diag}(\pi)$

Proof.

Property 4. P and S have the same (set of) eigen values

Property 5. $\lambda_1(S) = 1$ with $\begin{bmatrix} \sqrt{\pi_1} \\ \vdots \\ \sqrt{\pi_n} \end{bmatrix}$ as the eigen vector

such that

$$\begin{aligned} S &= U \Lambda U^T \\ &= \begin{bmatrix} \sqrt{\pi_1} \\ \vdots \\ \sqrt{\pi_n} \end{bmatrix} \begin{bmatrix} \sqrt{\pi_1} & \cdots & \sqrt{\pi_n} \end{bmatrix} + \begin{bmatrix} u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} \lambda_2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} u_2^T \\ \vdots \\ u_n^T \end{bmatrix} \end{aligned}$$

- **Proof.** of the spectral bound

$$
\begin{aligned}
2\,|(p^{(0)})^T P^t - \pi^T|_{\mathrm{TV}} &= \sum_i |((p^{(0)})^T P^t - \pi^T)_i| \\
&= \sum_i \frac{|((p^{(0)})^T P^t - \pi^T)_i|}{\pi_i^{1/2}} \pi_i^{1/2} \\
&\le \|((p^{(0)})^T P^t - \pi^T)\Pi^{-1/2}\|\,\|\pi^{1/2}\| && \text{[Cauchy-Schwar} \\
&= \|((p^{(0)})^T P^t - \pi^T P^t)\Pi^{-1/2}\| \\
&= \|(p^{(0)} - \pi)^T \Pi^{-1/2} S^t\| \\
&\color{red}{\le} \|(p^{(0)} - \pi)^T \Pi^{-1/2}\|\,|\lambda_2|^t && \text{[Spectral analysis]} \\
&\le (1 + \frac{1}{\sqrt{\pi_{\min}}})\,|\lambda_2|^t && \text{[Triangular ineq.]}
\end{aligned}
$$

$$
\|(p^{(0)} - \pi)^T \Pi^{-1/2}\| \le \underbrace{\|(\pi)^T \Pi^{-1/2}\|}_{=1} + \underbrace{\|p^{(0)}\|\,\|\Pi^{-1/2}\|_2}_{\le 1/\sqrt{\pi_{\min}}}
$$

$$\|(p^{(0)} - \pi)\Pi^{-1/2} S^t\| \;\leq\; \|(p^{(0)} - \pi)\Pi^{-1/2}\| \, |\lambda_2|^t$$

1. $(p^{(0)} - \pi)^T \Pi^{-1/2}$ is orthogonal to the first singular vector of S
   - recall $P = \Pi^{-1/2} S \Pi^{1/2}$
   - largest eigenvalue of P is one with left and right eigen vectors π and $\mathbb{1}$
   - let $\pi^{1/2} = \Pi^{1/2}\mathbb{1}$
   - $S\pi^{1/2} = \pi^{1/2}$, since $S\pi^{1/2} = \Pi^{1/2} P \Pi^{-1/2} \Pi^{1/2} \mathbb{1} = \Pi^{1/2}\mathbb{1}$
   - hence, $\pi^{1/2} = \Pi^{1/2}\mathbb{1}$ is the eigenvector corresponding to the largest eigen value of S which is also one

$$(p^{(0)} - \pi)^T \Pi^{-1/2} \cdot \Pi^{1/2} \mathbb{1} \;=\; 0$$

2. if a is orthogonal to the first singular left vector of S, then

$$\|a^T S^t\| \;\leq\; \|a\| \sigma_2(S)^t$$

   - eigen value decomposition: $S = U \Lambda U^T$, where $UU^T = U^T U = \mathbf{I}$
   - $S_1 \equiv U_1 \lambda_1 U_1^T$, and $a^T S^t \;=\; a^T (S - S_1)^t$
   - $\|a^T S^t\| \;=\; \|a^T (S - S_1)^t\| \leq \|a\| \|S - S_1\|_2^t = \lambda_2^t \|a\|$

the spectral properties of some simple random walks on graphs

- complete graph:

$$P = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \;, \quad \text{with } |\lambda_2| = 0 \;, \; T_{\rm mix} \propto \frac{1}{\log(1/0)}$$

- cycle:

$$P = \frac{1}{2}\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix} \;, \quad \text{with } |\lambda_2| = 1 - O(1/n^2) \;, \; T_{\rm mix} \propto n^2$$

- star:

$$P = \begin{bmatrix} 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \;, \quad \text{with } \lambda_2 = -1 \;, \; T_{\rm mix} = \infty$$

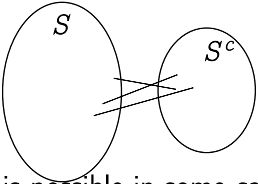
Bounding mixing time via conductance [Exercise 8.1]

- spectral analysis, and in particular the second largest eigen value of P, gives a means to bound the mixing time
- however, computing the spectral gap can be challenging
- **Cheeger's inequality** provides a bound on the spectral gap:

$$\frac{1}{1-\lambda_2} \leq \frac{2}{\Phi^2}$$

where **conductance** Φ of P is defined as

$$\Phi \triangleq \min_{S \subset \mathcal{X}^n} \frac{\sum_{x \in S, y \in S^c} \pi_x P_{xy}}{\pi(S)\pi(S^c)}$$



- direct computation of Φ is possible in some cases

$$T_{\text{mix}}(\epsilon) \leq \frac{2 \log \frac{2}{\epsilon \sqrt{\pi_{\min}}}}{\Phi^2}$$

Bounding mixing time via coupling

- **Define.** a **coupling** of two random variables X and Y with distributions $\mu_X(x)$ and $\mu_Y(y)$ is a construction of a joint probability distribution over (X, Y), i.e. $\mu(x, y)$ such that the marginals are preserved: $\sum_y \mu(x, y) = \mu_X(x)$ and $\sum_x \mu(x, y) = \mu_Y(y)$

- **example.** two (marginal) Gaussians $\mu(x) \sim \mathcal{N}(0, 1)$ and $\mu(y) \sim \mathcal{N}(0, 4)$
  - independent
  - Y=2X

- **example.** two (marginal) Bernoulli $X \sim \text{Bern}(p)$ and $Y \sim \text{Bern}(q)$
  - ⋆ independent
  - ⋆ construction from $U[0,1]$

- how closely can we couple X and Y?
  in other words, what is

$$\min_{\text{coupling of } \mu_x, \mu_Y} \mathbb{P}(X \neq Y)$$

- **Coupling lemma.** for two (continuous or discrete) random variables X and Y in the same domain,

$$\|\mu_X - \mu_Y\|_{\mathrm{TV}} = \min_{\text{couplings of } \mu_X, \mu_Y} \mathbb{P}(X \neq Y)$$

- **proof.**

$$\begin{aligned}
\mathbb{P}(X \neq Y) &= 1 - \sum_x \mu_{X,Y}(x,x) \\
&\geq \sum_x \Big\{ \mu_X(x) - \min\{\mu_X(x), \mu_Y(x)\} \Big\} \\
&= \sum_x \max\{0, \mu_X(x) - \mu_Y(x)\} \\
&= \frac{1}{2} \sum_x \big| \mu_X(x) - \mu_Y(x) \big|
\end{aligned}$$

further, exists $\mu(x,y)$ such that $\mu(x,x) = \min\{\mu_1(x), \mu_2(x)\}$, and $\mu(x,y) = \frac{(\mu_X(x) - \mu(x,x))(\mu_Y(y) - \mu(y,y))}{1 - \sum_x \mu(x,x)}$

- example of an optimal coupling

$$X = \begin{cases} 0 & \text{w.p. } p \\ 1 & \text{w.p. } 1-p \end{cases} \qquad Y = \begin{cases} 0 & \text{w.p. } q \\ 1 & \text{w.p. } 1-q \end{cases}$$

need to construct a probability distribution over X and Y

$\min\{p, q\}$	$\max\{0, p-q\}$	p
$\max\{0, q-p\}$	$\min\{1-p, 1-q\}$	$1-p$
q	$1-q$	

this naturally extends to larger alphabet. Equivalently, one could draw $Z \sim$ Uniform[0,1], then coupling is nothing but determining intervals in $[0,1]$ for each output of X and Y. For example, the optimal coupling is

$$X = \begin{cases} 0 & \text{if } Z \in [0,p] \\ 1 & \text{otherwise} \end{cases} \qquad Y = \begin{cases} 0 & \text{if } Z \in [0,q] \\ 1 & \text{otherwise} \end{cases}$$

- **Corollary of the coupling lemma.** total variation can be upper bounded by any coupling,

$$|\mu_X - \mu_Y|_{\mathrm{TV}} \;\leq\; \mathbb{P}_{(X,Y)}(X \neq Y)$$

Coupling for bounding $T_{\rm mix}$ of Gibbs sampling

- let X_t and Y_t be random states after t transitions according to P with initial state X_0 and Y_0
- **Corollary of the coupling lemma.** for any coupling of X_t and Y_t,

$$|\mu_{X_t} - \mu_{Y_t}|_{\rm TV} \;\leq\; \mathbb{P}_{(X_t,Y_t)}(X_t \neq Y_t)$$

- **Strategy.** to get a tight bound on the total variation, we need to construct good coupling.

$$\begin{aligned} |\mu_{X_t} - \pi|_{\rm TV} &\leq \max_{\mu_{X_0},\mu_{Y_0}} |\mu_{X_t} - \mu_{Y_t}|_{\rm TV} \\ &\leq \max_{\mu_{X_0},\mu_{Y_0}} \mathbb{P}(X_t \neq Y_t) \end{aligned}$$

we consider a particular coupling of two Gibbs sampling chains for $x, y \in \{0,1\}^n$

1. draw uniform $I \in [n]$
2. draw x'_I from $\mu(x'_I|x_{\partial I})$ and y'_I from $\mu(y'_I|y_{\partial I})$ using the optimal coupling

- Bounding $\mathbb{P}_{(X_t,Y_t)}(X_t \neq Y_t)$ by **path coupling**

[R. Bubley and M. Dyer, FOCS 1997]

  - **Define.** $D(x,y)$ is the minimal number of allowed moves in the transition matrix P to go from x to y (e.g. Hamming distance for Gibbs sampling)
  - **Idea.** if we can construct a coupling such that

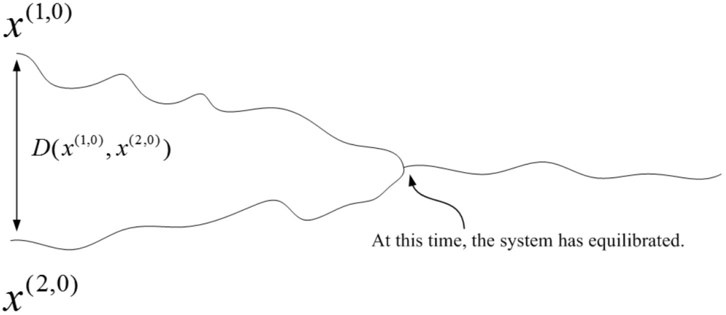
$$\mathbb{E}[D(x_{t+1},y_{t+1})|x_t,y_t] \leq \alpha D(x_t,y_t) \tag{1}$$

for some $0 < \alpha < 1$, then

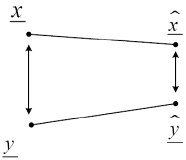
$$\begin{aligned} |\mu_{X_t} - \mu_{Y_t}|_{\mathrm{TV}} &\leq \mathbb{P}(X_t \neq Y_t) \\ &\leq \mathbb{E}[D(x_t,y_t)] \\ &\leq \alpha^t D(x_0,y_0) \\ \Rightarrow \quad T_{\mathrm{mix}}(\epsilon) &\leq \frac{\log \frac{D(x_0,y_0)}{\epsilon}}{\log \frac{1}{\alpha}} \end{aligned}$$

Path coupling for Gibbs sampling

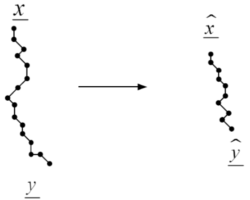
two Markov chains start at a distance as measured by $D(x^{(1,0)}, x^{(2,0)})$, and with the right coupling two sample path eventually converge and follow the same sample path after some (random) time



- **Path coupling.** to prove that $\mathbb{E}[D(x_{t+1}, y_{t+1}) | x_t, y_t] \leq \alpha D(x_t, y_t)$ it is sufficient to prove it for x_t and y_t that only differ in one vertex



Have to consider all possible pairs



Consider each step instead

Claim. If $\mathbb{E}[\, D(\hat{x}, \hat{y}) | D(x, y) = 1\,] \leq \alpha$ then Eq. (1) follows.
Proof sketch. consider a minimum length path from x to y:

$$p = (x, p_1, \ldots, p_{D(x,y)-1}, y)$$

which are, after one step of the Markov chain, mapped to

$$(\hat{x}, \hat{p}_1, \ldots, \hat{p}_{D(x,y)-1}, \hat{y})$$

by triangular inequality,

$$\begin{aligned} \mathbb{E}[D(\hat{x}, \hat{y}) | x, y] &\leq \mathbb{E}[D(\hat{x}, \hat{p}_1) + D(\hat{p}_1, \hat{p}_2) + \cdots + D(\hat{p}_{D(x,y)-1}, \hat{y})] \\ &\leq \alpha\, \mathbb{E}[D(x, y)] \end{aligned}$$

for some graphical models, path coupling constant α can be bounded, e.g.

$$\mu(x) \;=\; \frac{1}{Z}\exp\Big\{\sum_{i,j\in E}\theta_{ij}x_ix_j\Big\}$$

- **Claim.** for Gibbs sampling on **Ising models**,

$$\mathbb{E}[D(x_{t+1},y_{t+1})|D(x_t,y_t)=1] \;\le\; 1-\frac{1-d_{\max}\tanh(\theta_{\max})}{n}$$

- hence, Gibbs sampling mixes fast when $d_{\max}\tanh(\theta_{\max})<1$

- **Step 1. Construction of a good coupling.** to prove the claim, we consider a particular coupling of two Gibbs sampling chains
  1. draw uniform $I\in[n]$
  2. draw x'_I from $\mu(x'_I|x_{\partial I})$ and y'_I from $\mu(y'_I|y_{\partial I})$ coupled in the following way

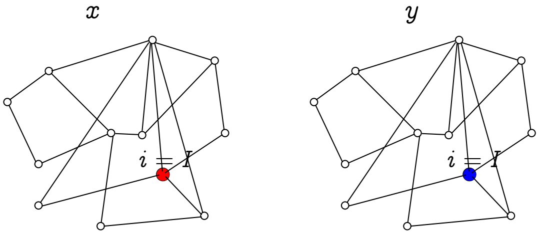
  2-1. draw a random $Z\sim$ Uniform$[0,1]$

  2-2. let

$$x'_I=\begin{cases}+1 & \text{if } Z\in[0,\mu(x'_I=+1|x_{\partial I})]\\ -1 & \text{otherwise}\end{cases} \qquad y'_I=\begin{cases}+1 & \text{if } Z\in[0,\mu(y'_I=+1|y_{\partial I})]\\ -1 & \text{otherwise}\end{cases}$$

- **Step 2. Analysis of the distance.** we are left to show that

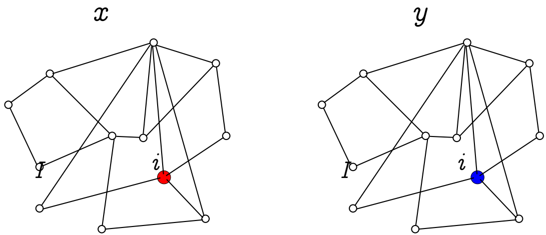
$$\mathbb{E}[D(x',y')|x \text{ and } y \text{ differ only at } i] \;\leq\; 1+\frac{1}{n}\Big\{-1+\sum_{j\in\partial i}|\tanh(\theta_{ij})|\Big\}$$



case 1. if $I = i$, $D(x',y')$ reduces to 0

$$\mathbb{E}[D(x',y')|x \text{ and } y \text{ differ only at } i, I=i] = 0$$

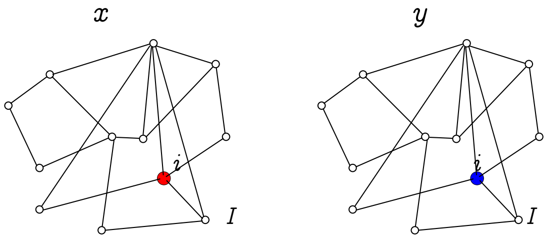
this happens with probability $1/n$



case 2. if $I \notin \{i\} \cup \partial i$, $D(x', y')$ remains at 1

$$\mathbb{E}[D(x', y')|x \text{ and } y \text{ differ only at } i, I \notin \{i\} \cup \partial i] = 1$$

this happens with probability $1 - \frac{1+|\partial i|}{n}$



case 3. if $I \in \partial i$, $D(x', y')$ can increase with probability

$$|\mu(x_I = +|x_{\partial I}) - \mu(y_I = +|y_{\partial I})| =$$

$$\left|\frac{A^{(+)}\psi_{iI}(+,+)}{A^{(+)}\psi_{iI}(+,+) + A^{(-)}\psi_{iI}(+,-)} - \frac{A^{(+)}\psi_{iI}(-,+)}{A^{(+)}\psi_{iI}(-,+) + A^{(-)}\psi_{iI}(-,-)}\right|$$

where $A^{(+)} = \prod_{j \in \partial I \setminus \{i\}} \psi_{jI}(x_j, +)$, and $A^{(-)} = \prod_{j \in \partial I \setminus \{i\}} \psi_{jI}(x_j, -)$

- **Claim.** for Ising model with $\psi(x_i, x_I) = e^{\theta_{iI} x_i x_I}$, the probability is bounded by $|\tanh(\theta_{iI})|$
- **proof.** in the case of $\theta_{iI} > 0$, we want to show that

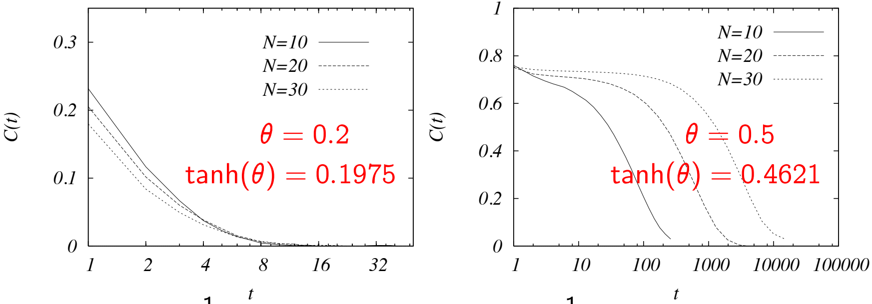
$$\begin{aligned}
& \frac{A^{(+)} e^{\theta_{iI}}}{A^{(+)} e^{\theta_{iI}} + A^{(-)} e^{-\theta_{iI}}} - \frac{A^{(+)} e^{-\theta_{iI}}}{A^{(+)} e^{-\theta_{iI}} + A^{(-)} e^{\theta_{iI}}} \\
= & \frac{A^{(+)} A^{(-)} (e^{2\theta_{iI}} - e^{-2\theta_{iI}})}{(A^{(+)})^2 + (A^{(-)})^2 + A^{(+)} A^{(-)} (e^{2\theta_{iI}} + e^{-2\theta_{iI}})} \\
= & \frac{(e^{2\theta_{iI}} - e^{-2\theta_{iI}})}{(A^{(+)})^2 + (A^{(-)})^2 + (e^{2\theta_{iI}} + e^{-2\theta_{iI}})} \\
\leq & \frac{(e^{2\theta_{iI}} - e^{-2\theta_{iI}})}{2 + (e^{2\theta_{iI}} + e^{-2\theta_{iI}})} \quad = \quad \tanh(\theta_{iI})
\end{aligned}$$

where we used the fact that $A^{(+)} A^{(-)} = 1$ and it also follows that $(A^{(+)})^2 + (A^{(-)})^2 \geq 2$.

For Ising model,

$$\mu_{G,\theta}(x) = \frac{1}{Z_G(\theta)} \exp\left\{\theta \sum_{(i,j)\in E} x_i x_j\right\}.$$

we showed that Gibbs sampling mixed fast if $\tanh(\theta_{\max})\,\mathrm{deg}_{\max} < 1$. Experiment with G uniformly random with N vertices and $2N$ edges (average degree 4).



$$C(t) = \frac{1}{|V|}\sum_{i\in V} x_i(0)x_i(t), \qquad t = \frac{1}{|V|}\,[\text{number of steps}]$$

- **theorem.** [Mossel, Sly, 2010] Assume $\theta_{ij} = \theta > 0$. Then the Glauber Markov chain mixes rapidly provided

$$(k-1)\,\tanh(\theta) \quad < \quad 1$$

- **theorem.** [Gerschenfeld, Montanari, FOCS 2007] Assume

$$(k-1)\,\tanh(\theta) \quad > \quad 1$$

  then there exists a sequence of k-regular graphs $G_n = ([n], E_n)$ for which the Glauber Markov chain mixes in time $\exp\{\Theta(n)\}$.

- Is $(k-1)\tanh(\theta) = 1$ fundamental?
  - Recall computation tree $T^{(t,i)}$ is formed from a graphical model by considering a root node x_i and a tree of all non-backtracking (non-reversing) paths for length t.
  - **Proposition.** Let $\nu_i(x_i)$ be the BP estimate after t iterations, $\nu^{(t)}_{i\to j}(x_i)$ be the BP message, and $\mu^{(t,i)}(x_i)$ be the marginal of the root x_i on the computation tree $T^{(t,i)}$, with some boundary conditions to be specified with the model. Then,

$$\nu_i^{(t_0+t_1)}(x_i) \quad = \quad \mu^{(t_1,i)}(x_i)$$

    with the boundary condition of the computation tree set to $\nu^{(t_0)}_{j\to k}(x_j)$ for a node x_j in the boundary with parent node x_k.
  - **Proof.** proof by induction.
  - **Corollary.** Let $\partial T^{(t,i)}$ denote the boundary nodes of the tree. If

$$\max_{x_{\partial T^{(t,i)}}, x'_{\partial T^{(t,i)}}} \left| \mu^{(t,i)}(x_i | x_{\partial T^{(t,i)}}) - \mu^{(t,i)}(x_i | x'_{\partial T^{(t,i)}}) \right|_{\mathrm{TV}} \quad \le \quad \delta(t) \,, \quad (2)$$

    then, for all $t_1, t_2 \ge t$,

$$\left| \nu_i^{(t_1)}(x_i) - \nu_i^{(t_2)}(x_i) \right| \le \delta(t) \,.$$

    In particular, if $\delta(t) \to 0$ as t grows, then BP converges.

- **Define.** $B_i(t)$ as the subgraph of G that includes all nodes at most distance t from node x_i.
- **Corollary.** If $B_i(t)$ is a tree, and Equation (2) holds, then

$$\Big| \underbrace{\mu(x_i)}_{\text{actual marginal}} - \underbrace{\nu_i^{(t)}(x_i)}_{\text{BP estimate}} \Big| \le \delta(t) \, .$$

  In particular, if g is the girth (the length of the shortest cycle) of G, then we have

$$\Big|\mu(x_i) - \nu_i(x_i)\Big| \le \delta((g-1)/2)$$

- **Proof.** observe that $\mu(x_i) = \sum_{x^{(t)}} \mu(x_i|x^{(t)})\mu(x^{(t)})$ where $x^{(t)}$ are the nodes at distance t from x_i.

- the condition (2) is known as **correlation decay** and we established that correlation decay implies convergence of BP in general graphs and correctness of BP on locally tree-like graphs, but checking condition (2) can be challenging

Dobrushin's uniqueness criterion

- Dobrushin's criterion measures the strengths of interactions, and provides a sufficient condition for Condition (2).
- **Define.** Influence of j on i as

$$C_{ij} \triangleq \max_{x,x' \text{ that only differ at } j} \left|\mu(x_i = \cdot|x_{V\setminus i}) - \mu(x_i = \cdot|x'_{V\setminus i})\right|_{\mathrm{TV}}$$

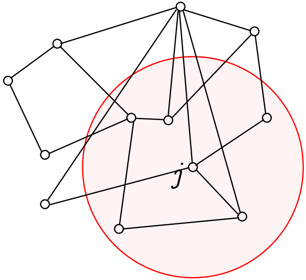
  - ★ $0 \leq C_{ij} \leq 1$
  - ★ $C_{ij} = 0$ unless $(i,j) \in E$
- **Theorem.**[Dobrushin, 1968] Small influence implies correlation decay. Let

$$\gamma \triangleq \max_{i \in V} \Big\{ \sum_{j \in \partial i} C_{ij} \Big\} \,.$$

Then,

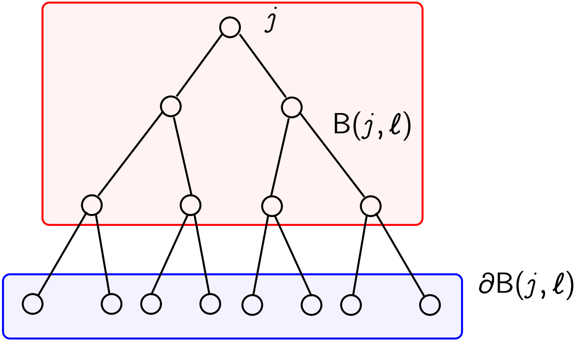
$$\max_{x,x'} \left|\mu(x_i = \cdot|x_{V\setminus B_i(t)}) - \mu(x_i = \cdot|x'_{V\setminus B_i(t)})\right|_{\mathrm{TV}} \leq \frac{\gamma^t}{1-\gamma}$$

Proof strategy



- bound influence on vertex j from those outside a ball of radius ℓ

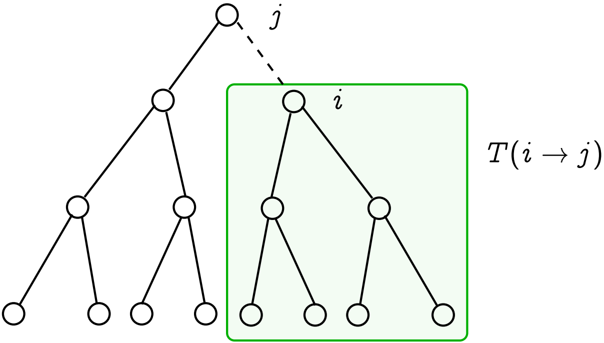
- assume neighborhood of j is a k-regular tree
- a graphical model satisfies **uniqueness condition** if



$$\sup_{y_{\partial B}, z_{\partial B}} \left| \mu(x_j | x_{\partial B} = y_{\partial B}) - \mu(x_j | x_{\partial B} = z_{\partial B}) \right| \leq \varepsilon(\ell) \downarrow 0$$

[In reality slightly stronger condition needed for proof]

Checking for uniqueness



$$h_{i\to j} \equiv \text{atanh}\mathbb{E}_{\mu,T(i\to j)}\{x_i\}\,.$$

Uniqueness: $h_{i\to j}$ asymptotically independent of boundary condition

Checking for uniqueness

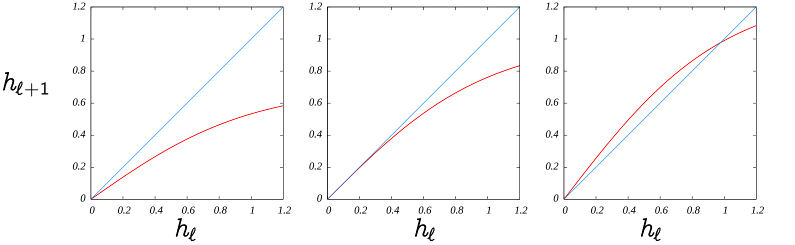
Exercise:

$$h_{i\to j} = \theta_i + \sum_{v\in \text{children}(i)} \text{atanh}\{\, \tanh\theta_{iv} \tanh h_{v\to i}\}\,.$$

- $\theta_{ij} = \beta$, $\theta_i = 0$,
- $x_{\partial \mathsf{B}(j,\ell)} = +1$, $x_{\partial \mathsf{B}(j,\ell)} = -1$ (monotonicity)

$$h_{\ell+1} = (k-1)\text{atanh}\{\, \tanh\beta \tanh h_\ell\}\,.$$

A one-dimensional recursion



$(k-1)\tanh\beta < 1$ $(k-1)\tanh\beta = 1$ $(k-1)\tanh\beta > 1$

- who cares about regular trees?
- regular trees are the worst case for decay of correlations

What about the lower bound?

Theorem (Gerschenfeld, Montanari, FOCS 2007)

Assume $(k-1)\tanh\beta > 1$.
Then there exists a sequence of k-regular graphs $G_n = (V_n = [n], E_n)$ for which the Glauber Markov chain mixes in time $\exp\{\Theta(n)\}$.

Proof.

Take G_n a uniformly random k-regular graph and prove that w.h.p.

$$\mathbb{P}_\mu\Big\{\sum_{i\in V} x_i = 0\Big\} = e^{-\Theta(n)},$$

$$\mathbb{P}_\mu\Big\{\sum_{i\in V} x_i > 0\Big\} = \mathbb{P}_\mu\Big\{\sum_{i\in V} x_i < 0\Big\} = \frac{1}{2} - e^{-\Theta(n)}.$$

Bottleneck! □

Are random graphs a curiosity?

No! Used as gadgets in

- Sly, *Computational transition at the uniqueness threshold*, 2010
- A. Sly, N. Sun, *The Computational Hardness of Counting in Two-Spin Models on d-Regular Graphs*, 2012
- A. Galanis, D. Stefankovic, and E. Vigoda, *Inapproximability of the partition function for the antiferromagnetic Ising and hard-core models*, 2012
- ...

Theorem

For antiferromagnetic Ising models $\theta_{ij} = -\theta < 0$, $\theta_i = 0$, *the partition function cannot be approximated unless RP=NP.*

$Q_n(\beta) \equiv \mathbb{P}_\mu\{ \sum_{i \in V} x_i = 0 \}$

$$\mu_{G,\beta}(x) = \frac{1}{Z_G(\beta)} \exp\Big\{ \beta \sum_{(i,j)\in E} x_i x_j \Big\}$$

$$Q_n(\beta) = \frac{Z_G^*(\beta)}{Z_G(\beta)}, \qquad Z_G^*(\beta) \equiv \sum_{x:\,\langle x,1\rangle = 0} e^{\beta \sum_{(i,j)\in E} x_i x_j}$$

- Upper bound $Z_G^*(\beta)$ by $n^{10}\mathbb{E}_G Z_G^*(\beta)$.
- Lower bound $Z_G(\beta)$ by ...

Estimating Z_G

Theorem (A.Dembo, A.Montanari, Ann. Appl. Prob. 2010)

Let $\{G_n = (V_n, E_n)\}_{n\geq 1}$ be a sequence of graphs that (i) Is uniformly sparse; (ii) Converges locally to a unimodular Galton-Watson tree. Let $Z_n(\beta, B)$ be the Ising model partition function with $\theta_{ij} = \beta$, $\theta_i = B$. Then

$$\lim_{n\to\infty} \frac{1}{n} \log Z_n(\beta, B) = \text{[explicit expression]}$$
$$= \text{[Bethe free energy]}$$